Audio-visual Affect Recognition through Multi-stream Fused HMM for HCI

Zhihong Zeng*, Jilin Tu, Brian Pianfetti, Ming Liu, Tong Zhang, Zhenqiu Zhang,
Thomas S. Huang and Stephen Levinson
University of Illinois at Urbana-Champaign
{zhzeng, jilintu, mingliu1, zzhang6, huang, sel}@ifp.uiuc.edu, bpianfet@uiuc.edu

Abstract

Advances in computer processing power and emerging algorithms are allowing new ways of envisioning Human Computer Interaction. This paper focuses on the development of a computing algorithm that uses audio and visual sensors to detect and track a user's affective state to aid computer decision making. Using our Multi-stream Fused Hidden Markov Model (MFHMM), we analyzed coupled audio and visual streams to detect 11 cognitive/emotive states. The MFHMM allows the building of an optimal connection among multiple streams according to the maximum entropy principle and the maximum mutual information criterion. Person-independent experimental results from 20 subjects in 660 sequences show that the MFHMM approach performs with an accuracy of 80.61% which outperforms face-only HMM, pitch-only HMM, energy-only HMM, and independent HMM fusion.

1. Introduction

Changes in a person's affective state play a significant role in perception and decision making during human to human interactions. This fact has inspired the research field of "affective computing" which aims at enabling computers to express and recognize affect [9]. Perhaps the most fundamental applications of affective computing would be in human-computer interaction where the computer could detect and track a user's affective states and initiate communications based on this knowledge, rather than simply responding to a user's commands.

Research presented in paper is part of an ongoing federally funded project (ITR) [20] which is to contribute to the development of a multimodal human-computer intelligent interaction (HCII) environment. The concepts and tools resulting from this project are applied to an educational context for evaluation. This education-based testbed focuses on the ability of the computing environment to help middle school students learn math and science concepts through exploration. This project uses a proactive computing learning environment to achieve two goals. The first goal was to keep the children actively engaged in the learning activity. The second goal was to support the exploration of math and science phenomena enabling the children to increase their knowledge. This is done through the recognition of changes in the children's affective states (e.g. interest, boredom, frustration and puzzlement) and applying appropriate tutoring strategies (e.g. encouragement, transition, guidance, and confirmation).

The psychological study [16] indicated that judging someone's affective states, people mainly rely on facial expressions and vocal intonations. Thus, affect recognition should inherently be the issue of multimodal analysis. In this paper, we present our efforts toward audio-visual affect recognition. With an HCII application in mind, we used 11 affective states which indicate a user's cognitive/emotive state. And we focus on person-independent affect recognition in which testing data and training data are from different subjects. Thus, the variation of the data is more significant and classification is more challenging than person-dependent recognition in which testing data and training data are from the same subject.

For integrating coupled audio and visual streams, we present the multi-stream fused hidden Markov model (MFHMM) which can build optimal connection among multiple streams according to the maximum entropy principle and a maximum mutual information criterion. This person-independent affect recognition approach was tested in 660 sequences based on 20 subjects with 11 HCII-related affect states. The experimental results show that the MFHMM approach performs with an accuracy of 80.61% which outperformed face-only HMM, pitch-only HMM, energy-only HMM, and independent HMM fusion which assumes independence among tightly coupled streams.

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 **IEEE**

2. Related work

Researchers from many different disciplines are interested in the possibility of automated affect analysis and recognition. Recent advances in computing power and multimedia technologies are facilitating efforts toward audio-visual affect recognition. According to [1], four papers [2-5] reported advances of bimodal affect recognition. In addition, there have been three papers of bimodal emotion recognition [6-7][18] recently published in 2004.

Three of these seven papers only did person-dependent affect recognition [5][7][18]. [6] did not give the number of tested subjects, and did not mentioned feature normalization which is crucial for person-independent test either. Thus, we think that [6] should belong to person-dependent recognition. Compared with the rest reports of person-independent bimodal affect recognition [2-4], the progress in this paper includes:

1) 11 affective states are analyzed, including 7 basic emotions and 4 cognitive states (puzzlement, interest, boredom, and frustration). [2-4] only analyzed 5-7 basic emotions.

2) 20 subjects are tested. The numbers of subjects in [2-4] are at most five.

3) Multi-stream Fused HMM is applied in audio-visual fusion. [3-4] applied rule-based methods for combining two modalities. [2] applied the single-modal method in a sequential manner for bimodal recognition.

3. Database

The datasets used in previous papers [2-4] were small in the number of subjects, and were not related directly to human computer interaction. To overcome these problems, a large-scale database was collected [19]. This database consists of performances of 7 basic emotions (happiness, sadness, fear, surprise, anger, disgust, and neutral), and 4 cognitive states (interest, boredom, puzzlement and frustration).

The 20 subjects (10 female and 10 males) in our database consist of graduate and undergraduate students from different disciplines. The first frames of the videos used in our experiment are shown in Figure 1. This set of videos contains subjects with a wide variability in physiognomy. Although the subjects displayed affect expressions on request, the subjects chose how to express each state. They were simply asked to display facial expressions and speak appropriate sentences. Each subject was required to repeat each state with speech three times. Therefore, for every affective state, there are 3*20=60 video sequences. And there are totally 60*11=660 sequences for 11 affective states. The time of every sequence ranged from 2-6 seconds.

Figure 1. The videos used in our experiment

During labeling, start and end points of each emotion expression were determined by speech energy which is easy to detect. Once these segments were defined, corresponding points of facial feature and pitch sequences were labeled.

4. Facial feature extraction

A tracking algorithm called Piecewise Bezier Volume Deformation (PBVD) tracking [10] is applied to extract facial features in our experiment.

This face tracker uses a 3D facial mesh model which is embedded in multiple Bezier volumes. The shape of the mesh can be changed with the movement of the control points in the Bezier volumes. That guarantees the surface patches to be continuous and smooth. In the first video frame (frontal view of a neutral facial expression), the 3-D facial mesh model is constructed by manual or automatic selection [8] of landmark facial feature points. Once the model was fitted, the tracker can track head motion and local deformations of the facial features. At the current stage, only local deformations of facial features are used for affect recognition. These deformations are measured in terms of magnitudes of 12 predefined motions of facial features, called Motion Units (MUs), which are shown in Figure 2. The outputs of the face tracker corresponding to 12 MUs are used as facial features for later affect recognition in our experiment.

We notice that the movements of facial features are related to both affective states and content of speech. Thus, smooth facial features are calculated by averaging facial features at consecutive frames to reduce the influence of speech on facial expression, based on the assumption that the influence of speech

on face features is temporary, and the influence of affect is relatively more persistent.

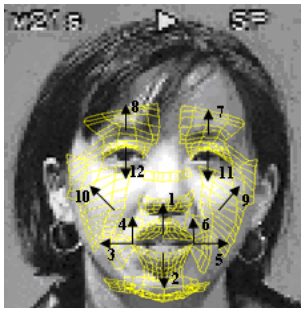

Figure 2. 12 facial Motion Units

Regarding person-independent affect recognition, facial feature normalization is crucial because every subject has different physiognomy. To express an affect, different subjects will display different magnitudes of 12 MUs. To overcome this difference, the neutral expression for each person has been used as the normalization standard. In detail, for a given subject, the magnitudes of 12 MUs at every frame were normalized by the corresponding feature means of the neutral expression of the same subject.

After the feature vector of each frame is normalized, it is quantized into 19-size codebook by vector quantization (VQ).

5. Audio feature extraction

For audio feature extraction, Entropic Signal Processing System named get_f0, a commercial software package, is used. It implements a fundamental frequency estimation algorithm using the normalized cross correlation function and dynamic programming [12]. The program can output the pitch F0 for fundamental frequency estimate, RMS energy for local root mean squared measurements, prob_voice for probability of voicing, and the peak normalized cross-correlation value that was used to determine the output F0. The experimental results in [13] showed pitch and energy are the most important factors in affect classification. Therefore, in our experiment, we only used these two audio features for affect recognition. Some prosody features, like frequency and duration of silence, could have implication in the HMM structure of energy and pitch.

Obviously, the emotional information in the voice depends on the subject and recording condition. The pitch varies widely from person to person. In general, males speak with a lower pitch than females. Thus, for a given subject, the pitch at every frame is normalized by the pitch mean of the neutral expression sequence of the same subject. The same is done for energy features to normalize amplitude change due to the speaker volume and the distance of a speaker for microphone.

Similarly to the visual feature quantization, the energy and pitch are quantized into 19-size codebook by vector quantization respectively.

6. Multi-stream fused HMM (MFHMM)

For integrating coupled audio and visual features, we propose multi-stream fused HMM (MFHMM) which constructs a new structure linking the multiple component HMMs which is optimal according to the maximum entropy principle and a maximum mutual information (MMI) criterion. MFHMM is a generalization of two-stream fused HMM [11]. It is suitable for the recognition problem which has more than two features. The advantages of MFHMM include: 1) every feature could be modeled by one component HMM so that the performance of every feature could be analyzed individually. And that analysis can be used for feature selection; 2) it reaches a better balance between model complexity and performance than other existing model fusion methods, like the coupled HMM (CHMM)[17] and mixed-memory HMM (MHMM) [15]. That is shown in [11]; 3) reliabilities of component HMM can be used to adjust the corresponding weights in final fusion. And if one component HMM fails due to some reason, the other HMM can still work. Thus, the final fusion performance can be robust.

In our affect recognition, we treat composite facial feature, speech energy and pitch as three tightly coupled streams. And three-stream fused HMM was used in our experiment.

6.1. Multi-stream Fused HMM

Consider n tightly coupled time series $\{O^{(i)}, i=1,\cdots,n\}$. Assume that the series $\{O^{(i)}, i=1,\cdots,n\}$ can be modeled respectively by n HMMs with hidden states $\{U^{(i)}, i=1,...,n\}$. In the fused HMM framework, an optimal solution for $p(O^{(1)};O^{(2)};\cdots;O^{(n)})$ according to the maximum entropy principle [11] is given by

$$\hat{p}(O^{(1)};O^{(2)};\cdots;O^{(n)}) = p(O^{(1)})p(O^{(2)})\cdots p(O^{(n)}) \bullet \frac{p(v^{(1)},v^{(2)},\cdots,v^{(n)})}{p(v^{(1)})p(v^{(2)})\cdots p(v^{(n)})} \quad (6.1.1)$$

The last term in the equation can be viewed as an enhancement/suppression factor, which absorbs some dependence among $\{O^{(i)}, i=1,\cdots,n\}$. The transforms

$$v^{(i)} = g_i(O^{(i)}) \quad i = 1,2,\cdots,n$$

were introduced so that $p(v^{(1)}, v^{(2)}, \cdots, v^{(n)})$ can more easily calculated than $p(O^{(1)}; O^{(2)}; \cdots; O^{(n)})$ and it reflects the statistical dependence among $\{O^{(i)}, i = 1, \cdots, n\}$.

According to the maximum mutual information (MMI) criterion, [11] proposed to fuse component HMMs together by connecting the hidden states of one HMM to the observation of another HMM. Thus n sets of transforms can be invoked with the *i*-th $(i = 1,2, \cdots, n)$ set of transform being

$$v^{(j)} = \begin{cases} U^{(j)} & j = i \\ O^{(j)} & j \neq i \end{cases} \qquad (j = 1,2, \cdots, n)$$

The *i*-th corresponding fusion model defined by (6.1.1) yields

$$\begin{aligned} &\hat{p}^{(i)}(O^{(1)}; O^{(2)}; \cdots; O^{(n)}) \\ &= p(O^{(1)}) p(O^{(2)}) \cdots p(O^{(n)}) \\ &\bullet \frac{p(U^{(i)}, O^{(1)}, \cdots, O^{(i-1)}, O^{(i+1)}, \cdots, O^{(n)})}{p(U^{(i)}) p(O^{(1)}) \cdots P(O^{(i-1)}) P(O^{(i+1)}) \cdots p(O^{(n)})} \\ &= p(O^{(i)}) p(O^{(1)}, \cdots, O^{(i-1)}, O^{(i+1)} \cdots, O^{(n)} \mid U^{(i)}) \end{aligned}$$

And assuming

$$\begin{aligned} &p(O^{(1)}, \cdots O^{(i-1)}, O^{(i+1)} \cdots, O^{(n)} \mid U^{(i)}) \\ &= \prod_{j \neq i, j=1}^{n} p(O^{(j)} \mid U^{(i)}) \end{aligned}$$

Although this conditional independence assumption is usually violated in practice, it has a good record in pattern recognition. The reason of the success of this assumption is attributed to the small number of parameters to be estimated. Some complicated algorithms without this assumption require more training data, and are more susceptible to local maximum during parameter estimation.

Thus, the estimate of $p(O^{(1)}; O^{(2)}; \cdots; O^{(n)})$ is given by

$$\begin{aligned} &\hat{p}^{(i)}(O^{(1)}; O^{(2)}; \cdots; O^{(n)}) \\ &= p(O^{(i)}) \prod_{j \neq i, j=1}^{n} p(O^{(j)} \mid U^{(i)}) \end{aligned} \qquad (6.1.2)$$

The structures defined by (6.1.2) are different for different i. (6.1.2) emphasizes the dependencies between hidden states $U^{(i)}$ and observations $\{O^{(j)}, j = 1, \cdots, n, j \neq i\}$, which requires that $U^{(i)}$ be reliably estimated. In practice, if the n component HMMs have different reliabilities, they may be combined by different weights to get a better result:

$$\begin{aligned} &\hat{p}(O^{(1)}; O^{(2)}; \cdots; O^{(n)}) \\ &= \sum_{i=1}^{n} \lambda^{(i)} \hat{p}^{(i)}(O^{(1)}; O^{(2)}; \cdots; O^{(n)}) \end{aligned} \qquad (6.1.3)$$

where $\sum_{i=1}^{n} \lambda^{(i)} = 1$

The weights of (6.1.3) could be set proportinal to the reliabilities of the component HMMs. The more reliable one component HMM is, the larger its weight. In our experiments, the performance of component HMMs increases from face-only, pitch-only and energy-only HMMs, so their corresponding weights are set to be 2.5, 3.5 and 4 respectively.

6.2. Learning Algorithm

The learning algorithm of the n-stream fused HMM includes three main steps.

1) n component HMMs are trained independently by the EM algorithm.
2) The best hidden state sequences of the component HMMs are estimated using the Viterbi algorithm
3) The coupling parameters between the n HMMs are estimated.

In step 1, the model parameters (the initial, transition, and observation probabilities) of individual HMMs are estimated. And step 2 infers hidden states $U^{(i)}$ $(i = 1,2, \cdots, n)$. The details of the EM and the Viterbi algorithms used for solving the above problems can be found in [14].

In step 3, the coupling parameters between the n HMMs are determined as follows:

$$B^{(i,j)} = \arg\max p(O^{(j)} \mid U^{(i)})$$
$$i, j = 1,2, \cdots, n, i \neq j$$

Since $O^{(i)}$ and $U^{(i)}$ $(i = 1,2, \cdots, n)$ are known, the above equations are a typical ML problem.

6.3. Inference Algorithm

In our application, inference is the process of computing $\hat{p}(O^{(1)}; O^{(2)}; \cdots; O^{(n)})$ in (6.1.3) given observation sequence $\{O^{(i)}, i = 1, \cdots, n\}$ and the model parameters corresponding to each affective state. And the affective state with maximum of $\hat{p}(O^{(1)}; O^{(2)}; \cdots; O^{(n)})$ is regarded as the recognition result.

According to (6.1.2), we first computer individually

$$\hat{p}^{(i)}(O^{(1)}; O^{(2)}; \cdots, O^{(n)})$$
$$i = 1,2, \cdots, n$$

Then their results are combined according to (6.1.3) to get

$$\hat{p}(O^{(1)}; O^{(2)}; \cdots; O^{(n)})$$

The individual inference algorithm of (6.1.2) is derived from the forward inference procedure of traditional HMM [14]. The only difference of our algorithm is that multiple stream observations instead of one-stream observation at time instants are taken

into account. In the other words, observation probability

$$p(O^{(i)} \mid U^{(i)}) \qquad (i = 1,2,\cdots,n)$$

in the forward-backward procedure [14] is replaced by

$$\prod_{j=1}^{n} p(O^{(j)} \mid U^{(i)}) \qquad (i = 1,2,\cdots,n)$$

7. Experimental results

The MFHMM person-independent affect recognition algorithm was tested on 20 subjects (10 females and 10 males). For this test, all of the sequences of one subject are used as the test sequences, and the sequences of the remaining 19 subjects are used as training sequences. The test is repeated 20 times, each time leaving a different person out (leave-one-out cross-validation). For every affective state, there are 3*20=60 expression sequences. Therefore, there are totally 60*11=660 sequences for 11 affective states. The time of every sequence is from 2-6 seconds.

In our experiment, the composite facial feature from video, energy and pitch features from audio are treated as three tightly coupled streams ($O^{(1)}, O^{(2)}, O^{(3)}$), and modeled by three component HMMs with 12 hidden states. We used the following five methods to make decisions and compared the recognition results:

1) face-only HMM;
2) pitch-only HMM;
3) energy-only HMM;
4) independent-HMM (IHMM): assuming $O^{(1)}, O^{(2)}$ and $O^{(3)}$ are independent, it combines the component HMMs by computing:

$$\hat{p}(O^{(1)}; O^{(2)}; O^{(3)}) = p(O^{(1)})p(O^{(2)})p(O^{(3)})$$

5) MFHMM: assuming statistical dependence among $O^{(1)}, O^{(2)}$ and $O^{(3)}$, it combines the component HMMs by computing (6.1.3) in which $\lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}$ are set to be 2.5, 3.5 and 4 individually.

In order to make our experiment comparable with the previous basic emotion recognition reports, we also did 7-state basic emotion recognition besides 11-state affect recognition. The affect recognition results in our experiment are shown in Table 1.

Table 1. Average accuracies of affect recognition

%	Face	Pitch	Energy	IHMM	FHMM
11 States	38.64	57.27	66.36	72.42	80.61
7 States	52.38	63.81	70.71	75.24	85.24

Among the five methods mentioned above, face-only HMM gave the poorest performance. The main reason is that speaking influences facial expressions. Especially, subjects seldom display expressive peaks which are main characteristic for pure facial expressions without speaking. Pitch-only and energy-only HMMs performed better than face-only HMM but worse than IHMM and MFHMM because both of IHMM and MFHMM combine information of face, pitch and energy which provide complementary information for recognition. IHMM gave worse performance than MFHMM because it assumes that $O^{(1)}, O^{(2)}$ and $O^{(3)}$ are independent. The performance of MFHMM is best on recognition rate, and the time of its training and inference is only a little more than IHMM.

Thus, the MFHMM reaches a good balance between model complexity (as well as computational complexity) and performance.

The more details of comparison of the five methods of 11-state affect recognition are presented in Table 2 which lists the recognition rate of each affect. In Table 2 for face-only HMM confusion matrix, the 3 cognitive states (frustration, interest and boredom) has the lowest recognition rates. That shows that it is difficult to judge these subtle cognitive states if only using information of facial expression. Fortunately, the audio features (pitch and energy) provide complementary information in these cognitive states which are shown in Table 2.

8. Conclusion

With an automatic affect recognizer, a computer can respond appropriately to the user's affective state rather than simply responding to user commands. In this way, the nature of the computer interactions would become more authentic, persuasive, and meaningful. This type of interaction is the ultimate goal of ITR project where attending to changes in the child's affective states leads to a high level of engagement and knowledge acquisition. To accomplish this end, this paper proposes a person-independent audio-visual affect recognition.

For integrating tightly coupled audio-visual streams, we applied the multi-stream fused HMM which is optimal according to the maximum entropy principle and the maximum mutual information criterion. Experimental results from analyzing 11 affect states of 20 subjects suggests that the MFHMM with the recognition rate of 80.61% outperformed face-only HMM, pitch-only HMM, energy-only HMM, and independent HMM fusion which assumes independence among tightly coupled streams.

Multimodal recognition of human affective states is a largely unexplored and challenging problem. The elicited nature of the affects performed in our database has the potential to differ from corresponding performances in natural settings. The next stage in the evaluation of this algorithm will be attempting to detect these affect states in human interactions where the states are performed naturally.

Acknowledgement

We like to thank Dr. Lawrence Chen for collecting the valuable data in this paper for audio-visual affect recognition. This work has been funded by National Science Foundation: Information Technology Research Grant# 0085980

References

[1] Pantic M., Rothkrantz, L.J.M., Toward an affect-sensitive multimodal human-computer interaction, Proceedings of the IEEE, Vol. 91, No. 9, Sept. 2003, 1370-1390

[2] Chen, L. and Huang, T. S., Emotional expressions in audiovisual human computer interaction, Int. Conf. on Multimedia & Expo 2000, 423-426

[3] Chen, L., Huang, T. S., Miyasato, T., and Nakatsu, R., Multimodal human emotion/expression recognition, Int. Conf. on Automatic Face & Gesture Recognition 1998, 396-401

[4] De Silva, L. C., and Ng, P. C., Bimodal emotion recognition, Int. Conf. on Automatic Face & Gesture Recognition 2000, 332-335

[5] Yoshitomi, Y., Kim, S., Kawano, T., and Kitazoe, T., Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face, in Proc. ROMAN 2000, 178-183

[6] Song, M., Bu, J., Chen, C., Li, N., Audio-Visual Based Emotion Recognition-A New Approach, CVPR 2004.

[7] Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S., Analysis of Emotion Recognition Using Facial Expressions Speech and Multimodal Information, ICMI 2004

[8] Tu, J., Zhang, Z., Zeng, Z. and Huang, T.S., Face Localization via Hierarchical Condensation with Fisher Boosting Feature Selection, In Proc. Computer Vision and Pattern Recognition, 2004.

[9] Picard, R.W., Affective Computing, MIT Press, Cambridge, 1997.

[10] Tao, H. and Huang, T.S., Explanation-based facial motion tracking using a piecewise Bezier volume deformation mode ,CVPR'99, vol.1, pp. 611-617, 1999

[11] Pan, H., Levinson S., Huang, T.S., and Liang, Z.P., A fused Hidden Markov Model With Application to Bimodal Speech Processing, IEEE Transaction on Signal Processing, Vol.52, No.3, 573-581, March 2004

[12] Talkin, D., A Robust Algorithm for Pitch Tracking, in Speech Coding and Synthesis, Kkeijn, W.B., and Paliwal, K.K., Eds., Amsterdam: Elsevier Science, 1995

[13] Kwon, O.W., Chan, K., Hao, J., Lee, T.W, Emotion Recognition by Speech Signals, EUROSPEECH 2003.

[14] Rabiner, L.R., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of IEEE, Vol.77, No.2, February, 1989

[15] Saul, L.k. and Jordan, M.I., Mixed memory Markov model: Decomposing complex stochastic processes as mixture of simpler ones, Machine Learning, Vol.37, 75-88, Oct. 1999

[16] Mehrabian, A., Communication without words, Psychol. Today, vol.2, no.4, 53-56, 1968

[17] Brand, M. and Oliver, N., Coupled hidden Markov models for complex action recognition, In Proc. Computer Vision Pattern Recognition, 201-206, 1997

[18] Zeng, Z., Tu, J., Liu, M., Zhang, T., Rizzolo, N., Zhang, Z., Huang, T.S., Roth, D., and Levinson, S., Bimodal HCI-related Affect Recognition, ICMI 2004

[19] Chen, L.S, Joint Processing of Audio-Visual Information for the Recognition of Emotional Expressions in Human-Computer Interaction, PhD thesis, UIUC, 2000

[20] itr.beckman.uiuc.edu

Table 2. Comparison of face-only HMM, pitch-only HMM, energy-only HMM, IHMM and MFHMM for 11 state affect recognition

(%)	neutral	happy	sad	angry	disgust	surprise	fear	frustrated	puzzle	interest	bore
face	0.45	0.40	0.32	0.58	0.43	0.43	0.52	0.18	0.42	0.25	0.27
pitch	0.92	0.42	0.55	0.63	0.68	0.42	0.62	0.35	0.72	0.62	0.38
energy	0.92	0.53	0.33	0.67	0.85	0.60	0.72	0.63	0.78	0.58	0.68
IHMM	0.97	0.53	0.67	0.77	0.90	0.57	0.72	0.58	0.82	0.82	0.63
MFHMM	0.98	0.70	0.68	0.82	0.88	0.78	0.78	0.75	0.85	0.85	0.78